Alucinação da inteligência artificial

Por **BARBARA COELHO NEVES***

O fenômeno da alucinação na inteligência artificial revela o abismo entre a probabilidade estatística e a verdade factual, expondo os riscos de sistemas que priorizam a coerência semântica em detrimento da precisão

O que é a "alucinação da Inteligência artificial"?

É quando modelos generativos geram informações factualmente incorretas, mas com elementos e construção semântica que transmitem autoconfiança. De acordo com Spinak (2023), os modelos de Inteligência artificial generativa funcionam prevendo a próxima palavra com base no contexto anterior. Eles não sabem das coisas. Por isso, tendem a gerar declarações que parecem plausíveis, mas que, na verdade, são incorretas. Este fenômeno é conhecido como "alucinação" da Inteligência artificial.

A alucinação de inteligência artificial se refere a casos em que modelos de Inteligência artificial, particularmente os de grande escala, geram resultados incorretos ou sem sentido. A alucinação de Inteligência artificial ocorre quando uma Inteligência artificial fornece respostas que parecem corretas, mas estão erradas ou são inventadas (Relatório de Alucinação de IA, 2025). É como quando o ChatGPT ou o Gemini dizem algo com confiança, mas o conteúdo é falso. Esses erros geralmente são difíceis de detectar.

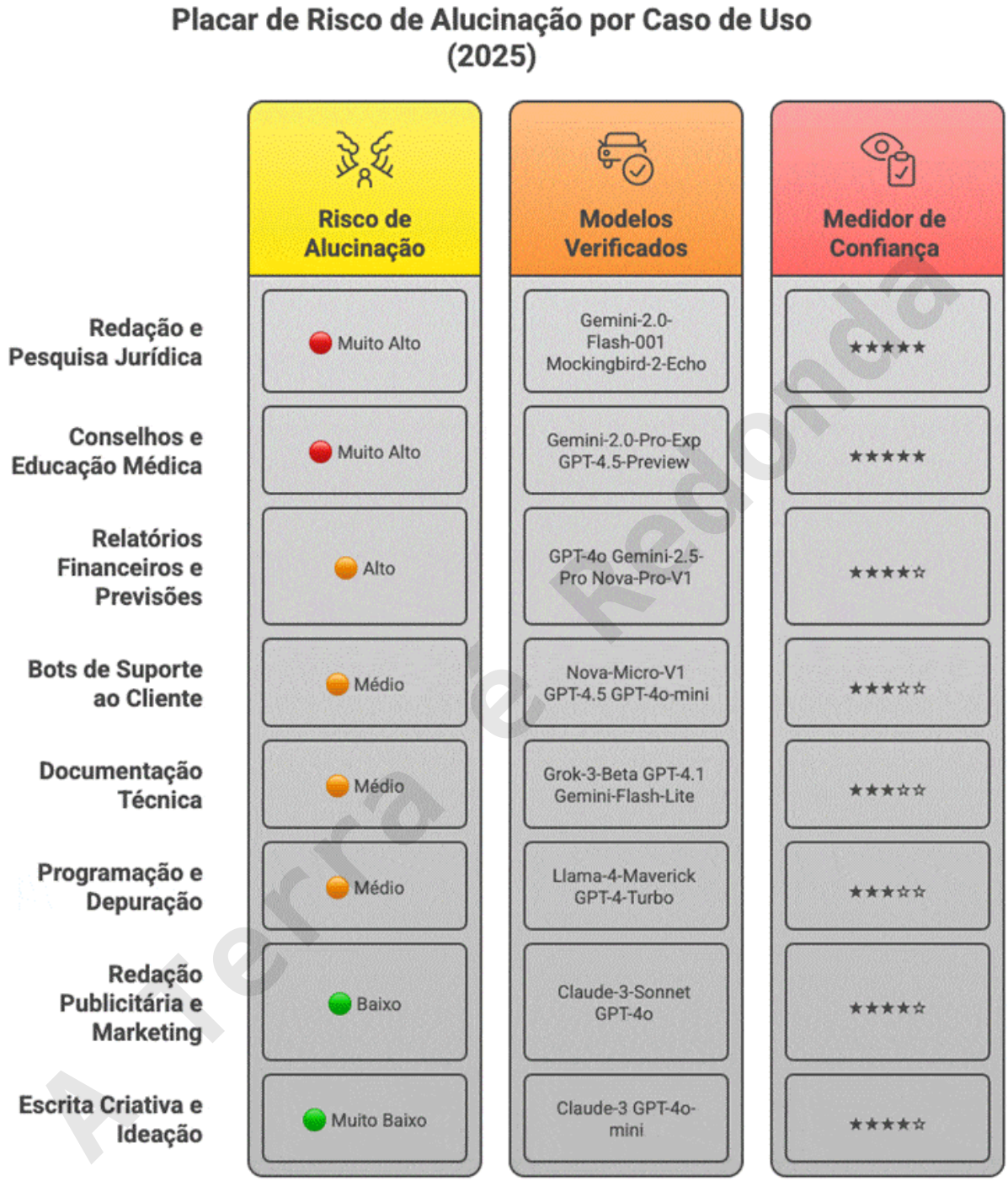
Os modelos com mais atualizações pagas podem reduzir estas preocupações, mas as versões de uso livre apresentam muitos problemas de "alucinação". As alucinações da Inteligência artificial acontecem porque os modelos de linguagem se baseiam em padrões probabilísticos com grande volume e variedade de dados, e não na compreensão factual ou na recuperação factual direta de dados verificáveis.

Alucinações são mais frequentes em modelos mais novos

Lançado em abril de 2025, o modelo o4-mini da OpenAI, que, apesar de apresentar avanços em tarefas como programação e matemática, demonstrou um aumento expressivo em alucinações, quando o sistema gera informações falsas.

Conforme descrito no trabalho do *International journal of science and research* publicado em janeiro de 2025, as alucinações podem acontecer por fatores intrínsecos quando os modelos de Inteligência artificial podem interpretar mal os dados devido a limitações em seus conjuntos de dados de treinamento, levando a resultados errôneos e por fatores extrínsecos quando os modelos recebem influências externas, como solicitações do usuário ou mal-entendidos contextuais, também podem contribuir para alucinações.

A seguir, a figura mostra o risco de cada modelo, considerando o uso.



Fonte: Elaboração própria com base na *Columbia Journalism Review.*

De acordo com o placar de Alucinação da Inteligência artificial de 2025 do *Columbia Journalism Review*, o Google Gemini-2.0-Flash-001 é o modelo de Inteligência artificial mais preciso com taxa de alucinação de apenas 0.7 por cento. Ele é seguido pelo Gemini-2.0-Pro-Exp e pelo OpenAI o3-mini-high com 0.8 por cento (Vectara, 2025).

Fonte: Coluna Tecnologia do *Brasil de Fato*.

As principais ferramentas que lideram na identificação de respostas alucinadas de grandes modelos de linguagem são: TruthfulQA, GPTZero, FactScore, Avaliação com Recuperação do Google (RAE) e RealityCheck.

Os pontos positivos destes recursos ajudam a verificar o conteúdo gerado por IA antes da publicação e funcionam em GPT-5, Claude, Gemini, dentre outros modelos LLMs. Vale salientar que funcionam por meio de APIs e permitem ajuste de rigor na detecção de alucinação. Contudo, tais ferramentas no momento do teste de detecção podem falhar com prompts abstratos ou criativos, sem falar que na versão gratuita apontam resultados de verificação com pouca precisão e podem exigir licenças pagas.

Um estudo da ChatBot Arena, uma plataforma de referência amplamente utilizada, descobriu que grandes empresas como OpenAI, Meta e Google testam muitas variantes de seus modelos de forma privada e liberam as pontuações de apenas as versões com melhor desempenho, deixando de fora resultados ruins, aponta um artigo do *Columbia Jornalism Review*. Os autores do estudo argumentam que esse processo deturpa as capacidades reais dos modelos.

O que tem a ver as alucinações da Inteligência artificial com Lacan

Existe uma analogia estrutural entre a alucinação com Inteligência artificial e o conceito de alucinação de Jacques Lacan na teoria psicanalítica. Uma pesquisa publicada em 2025 no Journal *"AI & Future Society (AFS)Technology, Ethics, and Governance in an Interdisciplinary Perspective*" sugere que interpretar a alucinação da Inteligência artificial através de uma lente lacaniana pode melhorar a compreensão desse fenômeno.

Se a teoria da psicose de Jacques Lacan revela a alucinação como um "produto necessário" que surge de lacunas na ordem simbólica, então o fenômeno da alucinação da Inteligência artificial também pode ser explicado dentro dessa estrutura. Ao contrário do sujeito psicótico, a Inteligência artificial não possui uma estrutura inconsciente; pois ela não é movida pelo desejo.

A rede simbólica na qual se baseia é sustentada por distribuições de probabilidade, superficialmente muito coerentes, mas carentes de um significante final. Em outras palavras, o modelo prioriza a captura de correlações entre palavras em vez de realmente compreender as relações causais no mundo real. Pode-se dizer que a "ordem simbólica" do modelo é inerentemente incompleta. O modelo não consegue distinguir fato de ficção por meio de um mecanismo autossuficiente, nem estabelecer um sistema restrito para verificação externa. Portanto, quando encontra lacunas de informação durante o

raciocínio ou a geração, ele não simplesmente interrompe a operação. Em vez disso, assim como o sujeito psicótico diante de Jacques Lacan, ele é compelido a compensar o vazio na ordem simbólica com alucinações.

A tarefa de um modelo de linguagem é essencialmente prever a próxima palavra por meio de aprendizado estatístico, em vez de operar dentro de uma "ordem da verdade" estável (Wang, 2025). Isso também explica por que as alucinações da Inteligência artificial frequentemente se manifestam como narrativas que possuem certeza e auto consistência, produzindo conteúdo que se conforma à lógica semântica, mas não à lógica factual. Essa situação não é produto de um erro acidental, mas de uma geração estrutural que ocorre necessariamente na ausência de ancoragem factual.

O fenômeno da "alucinação de Inteligência artificial" é um problema significativo que afeta a credibilidade dos sistemas de Inteligência artificial. Desse modo, a "alucinação da Inteligência artificial" também agrava a situação da desinformação, pois se refere à geração de informações distorcidas ou fictícias. Esse tipo de informação, muitas vezes convincente, mas incorreta, pode incluir erros lógicos, distorções de fatos e invenções que não são fundamentadas ou baseadas em evidências verdadeiras.

Desse modo, os tipos de alucinação são: factuais (dados, datas e eventos inexistentes), atribuição (citações falsas e fontes inventadas), lógicas (conclusões que não seguem as premissas).

Esse fenômeno apresenta desafios significativos em várias aplicações, incluindo os campos da saúde e pesquisa, onde imprecisões podem levar à desinformação e comprometer a tomada de decisões.

Isso também explica por que as alucinações da Inteligência artificial frequentemente se manifestam como narrativas que possuem certeza e autoconsistência, produzindo conteúdo que se conforma à lógica semântica, mas não à lógica factual. Essa situação não é produto de erro acidental, mas de uma geração estrutural que ocorre necessariamente na ausência de ancoragem factual. Como Lacan apontou, a razão pela qual as alucinações psicóticas são profundamente acreditadas pelo sujeito é precisamente porque elas não dependem de mediação simbólica, mas são impostas à experiência diretamente de maneira "real".

Supor que a saída de conteúdo dos modelos de Inteligência artificial seja precisa, sem verificação, pode levar a muitos resultados negativos, como: (i) danos de reputação; (ii) perdas financeiras; (iii) exposição legal.

Assim, vale a pena refletir com base em Wang (2025) que conclui que se a alucinação psicótica na teoria lacaniana revela a inevitabilidade estrutural da fissura entre o sujeito e a ordem simbólica, então a alucinação da Inteligência artificial, em outro nível, nos força a refletir sobre a relação entre linguagem, verdade e conhecimento.

***Barbara Coelho Neves** *é professora do Instituto de Ciência da Informação da UFBA. Autora, entre outros livros, de* Tecnologia e mediação (*Editora CRV*). [https://amzn.to/3H6GBf2]

Referências

SPINAK, E. IA: Como detectar textos produzidos por *chatbox* e seus plágios. Scielo em Perspectiva, November 17, 2023 14:30. https://blog.scielo.org/blog/2023/11/17/ia-como-detectar-textos-produzidos-por-chatbox-e-seus-plagios/

Novos modelos de IA da OpenAI têm mais 'alucinações' que os anteriores; entenda, *Exame*, 2025. https://exame.com/inteligencia-artificial/novos-modelos-de-ia-da-openai-tem-mais-alucinacoes-que-os-anteriores-entenda/?utm_source=copiaecola&utm_medium=compartilhamento.

WANG, Y. A Lacanian Interpretation of Artificial Intelligence Hallucination. *AI & Future Society*, *1*(1), 2025,

13-16.https://doi.org/10.63802/afs.v1.i1.93

VECTARA. Relatório de Alucinação de IA 2025: Qual IA Alucina Mais?, 2025. https://www.allaboutai.com/pt-br/recursos/alucinacao-em-llms/#which-llm-hallucinates-the-most

COLUMBIA Journalism Review. Journalists Need Their Own Benchmark Tests for AI Tools: The performance tests used by AI companies don't measure what matters in the newsroom, September 18, 2025.